Classification of Paper-Based Archival Records Using Neural Networks

Jussara Teixeira
Institute of Information and Communication Technology
PRODEST
Espírito Santo, Brazil
0000-0002-6152-0291

Juliana Almeida
State Public Archives
APEES
Espírito Santo, Brazil
juliana.almeida@ape.es.gov.br

Tânia Gava
Federal University
UFES
Espírito Santo, Brazil
tania.gava@ufes.br

Raphael Lugon Campo Dall'Orto
Institute of Information and Communication Technology
PRODEST
Espírito Santo, Brazil
raphael.dallorto@prodest.es.gov.br

José Márcio Moraes Dorigueto
Institute of Information and Communication Technology
PRODEST
Espírito Santo, Brazil
jose.dorigueto@prodest.es.gov.br

***Abstract*—This paper presents the application of Artificial Intelligence (AI) techniques to support the classification of paper-based archival records managed in the Electronic Process System (SEP) of the State of Espírito Santo, Brazil. Originally implemented in 1986 on a mainframe platform and migrated to a web-based environment in 2010, SEP contained more than 4.3 million unclassified records. To address this backlog, different supervised learning algorithms were evaluated, including Neural Networks, Decision Tree, Random Forest, and SGD Classifier, using TF–IDF vectorization. The experiments compared training sets of 20,000 and 200,000 examples and a test set consisting of 50,208 manually classified records. Results demonstrated that the Multilayer Perceptron (Neural Network) achieved an accuracy of 97.12% with low computational cost, outperforming traditional classifiers in scalability and adaptability. The implementation of a modular, container-based machine learning stack enabled large-scale automation, supporting the classification of more than 1.2 million records in 2025, with human supervision in specific stages of the process. During model application, clusters of processes with similar textual and contextual characteristics were identified, allowing a single classification to be assigned to entire groups, thus reducing manual effort and increasing consistency. This work contributes to the field of Computational Archival Science by demonstrating how AI can enhance functional classification while preserving archival principles of provenance and organicity.**

***Index Terms*—Archival Science, Artificial Intelligence, Computational Archival Science, Neural Network, Record Classification**

I. Introduction

The Electronic Process System (SEP) was launched in 1986 to modernize administrative workflows within the public administration of the state of Espírito Santo. However, over time, a large volume of archival records accumulated without proper classification. This phenomenon is not unique to Brazil: several electronic record management systems face difficulties in fully integrating archival principles, resulting in masses of unclassified records that compromise access, preservation, and institutional transparency [1].

In the specific case of SEP system, the coexistence of structured processes and a backlog of accumulated records exposed the urgency for scalable solutions capable of articulating archival requirements with advanced computational techniques. In this scenario, the use of Artificial Intelligence (AI) methods, especially neural networks, emerges as a promising alternative to automate classification and reduce the untreated collection. In addition to meeting an immediate administrative need, this initiative is connected to the international agenda of Computational Archival Science (CAS), which seeks to bring archival theory and computational methods closer together with a view to expanding the organization, access, and preservation of records.

The initial functionality of the SEP system, which uses AI to support end users in selecting the appropriate classification for administrative records, as well as the archival context of the Executive Branch of Espírito Santo and the preliminary results of that deployment, were presented in Teixeira *et al.* [2]. The present article expands that work by detailing the methodological steps involved in designing the classifier, presenting the experiments conducted during model evaluation, and outlining the machine learning components developed to support large-scale use.

To fully understand the results achieved, it is necessary to contextualize the situation prior to the adoption of AI techniques in the SEP system. The Electronic Process System (SEP), originally implemented in 1986, operated on a mainframe platform. In 2010, the system underwent a significant technological migration and was converted into a web-based environment. During this transition, 4,383,917 unclassified records from the previous database were incorporated.

In the first years after the migration, the user agencies relied exclusively on the State Government's General Classification Scheme. Over time, they developed their own classification schemes to meet specific administrative needs, resulting in 633,002 classified processes by 2022.

In 2023, a coordinated institutional effort enabled the manual classification of an additional 1,870,015 records, bringing the total to 2,503,017. However, the heterogeneity of the content and the absence of consistent patterns made this task increasingly difficult, highlighting the need for new approaches.

II. Background

A. Archival science

Archival science is structured around fundamental principles such as provenance, *respect des fonds*, authenticity, and organicity. These concepts are essential to ensure the contextualization and reliability of archival records. Schellenberg [3] emphasizes that classification is one of the central functions of archival work, as it establishes the link between the records and the activities of the producing institution. Duchein [4] reinforces organicity as a principle according to which records constitute natural and functional sets, resulting from the activities of the producing entity. Duranti [5] emphasizes authenticity and diplomatics as indispensable theoretical tools for understanding the nature of records in contemporary environments.

B. Artificial Intelligence

The field of AI has evolved significantly, with an emphasis on machine learning and, in particular, neural networks. Russell and Norvig [6] present AI as a field of study that combines techniques of knowledge representation, reasoning, and learning to solve complex problems. Goodfellow *et al.* [7], when discussing deep learning, demonstrate the potential of neural networks in tasks such as text classification, pattern analysis, and natural language recognition, fundamentals that prove useful when applied to the archival science domain.

C. Artificial Intelligence applied to archival records

The application of computational methods to the archival science domain has been the subject of increasing reflection, especially in view of the increase in the volume and complexity of records produced in digital environments. Foscarini [8] emphasizes that archival classification must be based on a functional understanding of the activities of organizations, since it is the relationship between functions and records that ensures the organicity and contextualization of the collection. This functional perspective is essential when introducing automation and artificial intelligence (AI) technologies into the classification process, since such systems need to reflect the administrative structure and functional links that give meaning to the records.

Marciano *et al.* [9] point out that, in recent decades, archivists have used theories and methods originally designed for paper records, while the production and use of digital records follow different technical and social logics. The authors therefore propose the formalization of a new transdisciplinary field called Computational Archival Science (CAS), which combines archival thinking and computational thinking. This integration seeks to apply methods from data science, machine learning, computational linguistics, and graph analysis to support archival functions such as appraisal, description, preservation, and access. CAS, as defined by the authors, aims to employ computational resources for the large-scale processing, analysis, and preservation of records and archives, without breaking with the archival principles of provenance, authenticity, and reliability.

In the context of public administration, Pareja [10] highlights that AI has established itself as a driver of digital transformation, especially through process automation and improved government efficiency. The author notes that AI applications in government—such as automatic information classification, predictive analysis, and intelligent data processing—require adequate technological infrastructure, governance policies, and institutional training. These same conditions are equally necessary for the use of AI in public archives, where automation must be accompanied by responsible management and transparency practices.

In this sense, the contributions of Foscarini [8], Marciano *et al.* [9], and Pareja [10] converge in indicating that the use of AI in archival science must respect the functional structure of the collections, the reliability of the records, and the governance of the technologies involved. The integration of archival principles and computational methods offers the potential to increase the efficiency of processes and strengthen preservation and access, provided that the focus remains on maintaining the context and authenticity of the records.

III. Methodology

The journey to build an Artificial Intelligence system for Classifying Paper-Based Processes began with internal discussions at the Espírito Santo Institute of Information and Communication Technology (PRODEST) about possible approaches to solving the problem at hand. These initial discussions involved several teams and experts who analyzed different methods and technologies that could be employed. As a result of these discussions, an interest group dedicated to the project was formed, and the search for external partners to conduct a Proof of Concept (PoC) began.

A PoC is an important step in the development of new projects, as it allows concepts to be tested on a smaller scale before committing significant resources to a full implementation. For the PoC of this project, five companies were invited to participate, but only one of them managed to complete the entire process. The result delivered by the company was evaluated in detail, but it still needed adjustments and improvements to fully meet the needs of the problem solution.

At the same time, the group formed at PRODEST began conducting its first experiments with AI. The goal was not only to find a technological solution to the problem, but also to initiate AI studies and implementations within PRODEST, fostering an environment of innovation and continuous learning. Based on these experiments, it was possible to develop and deliver an MVP (Minimum Viable Product) that demonstrated the feasibility of using AI for process classification.

With the MVP proving to be a viable solution, the team moved forward with the publication of the first version of the record classification API, integrated with the SEP system. This API allowed the SEP system to use AI technology in an integrated and efficient manner.

IV. Algorithm Selection and Model

The process of algorithm selection and model design was conducted systematically, taking into account both performance aspects and computational limitations. Several supervised classification techniques were evaluated, including Artificial Neural Networks (MLP), Decision Tree Classifier, Random Forest Classifier, and SGD Classifier. All approaches were applied to the same dataset after preprocessing using TF–IDF (Term Frequency–Inverse Document Frequency) as a text vectorization technique [11].

The experiments considered training sets with 20,000 and 200,000 examples, in addition to an independent set of 50,208 examples for manual testing. Table I summarizes the results obtained by each model, presenting accuracy, training time, and a detailed analysis of the errors made, classified according to different causes.

Among the models evaluated, the SGD Classifier achieved the lowest performance, with accuracies of 86.96% (20K) and 93.04% (200K), and required more than 1 h 27 min of training with 200,000 examples—a high cost considering its inferior performance. This result reinforces the sensitivity of SGD to hyperparameters and data scale [12].

Tree-based models achieved more satisfactory results. The Decision Tree Classifier reached 99.05% accuracy with 20,000 examples and 99.72% with 200,000. However, its training time increased disproportionately from 34s (20K) to more than 21 min (200K), indicating nonlinear scalability issues. The Random Forest model obtained accuracies of 98.90% (20K) and 99.53% (200K), with training times of 39s and 15 min 57s, respectively. Both models demonstrated good classification capability, but with considerable time growth as the dataset increased.

The neural network used followed the Multilayer Perceptron (MLP) architecture, implemented with the Keras library [13]. It featured one hidden layer with ReLU activation and one output layer with sigmoid activation, optimized using the Adam algorithm [14] and the categorical cross-entropy loss function. Although the code included Dropout as a regularization mechanism, this functionality was disabled in the main experiments [15].

The initially tested neural network, composed of 5 neurons and 10 training epochs, achieved 97.06% accuracy with 20,000 examples and 97.12% with 200,000 examples, with training times of 45s and 8 min 16s, respectively. Although its initial performance was slightly lower than that of the tree-based models, it showed a good balance between performance and computational cost. Furthermore, the errors made by the neural network were qualitatively analyzed in Table I, being mainly attributed to the low frequency of some examples in the training set and the occurrence of unknown words (13.64%–35.46% of the errors, depending on the model and scenario).

To further explore the neural network's potential, additional tests were carried out with more robust configurations: Neural Network A (10 neurons, 20 epochs) and Neural Network B (60 neurons, 30 epochs), as shown in Table II. These models achieved higher accuracies (99.42% and 99.82%, respectively), but with significantly longer training times (up to 43 min).

Despite the accuracy gains, the final model selected was the neural network from Table I (5 neurons, 10 epochs), which, although simpler, offered an excellent balance between accuracy, training time, and scalability. This choice considered the practical needs of the application scenario, prioritizing an efficient and easily replicable solution, without excluding the possibility of employing more complex architectures in future system versions.

V. Results

In 2024, Machine Learning (ML) techniques were introduced, initially as a decision-support system that required user confirmation. This pilot phase enabled the classification of 78,359 processes and established the methodological foundation for large-scale automation implemented in 2025. At that stage, SEP system achieved 1,246,869 new automatic classifications, with human supervision limited to validation and exception handling. Currently, the system manages 8,671,547 records, of which 3,828,245 have already been classified, allowing for their proper final disposition.

As shown in Figure 1, the number of classified records increased significantly after the adoption of machine learning techniques in 2024. During the large-scale execution phase, the analysis of classification patterns allowed the identification of clusters of similar processes. This grouping enabled the propagation of a single classification to entire sets, improving consistency and reducing manual validation efforts.

VI. Reusability of the Approach

To facilitate the development of classification and regression projects, a modular *machine learning* stack was built and published in a public repository, allowing any researcher or developer to download and use the infrastructure. The stack supports different models and processing flows, providing integration between critical components such as the database, message queues, API, processing workers, and model registry. The full implementation is available at https://github.com/prodest/prodest-ml-stack.

The implementation of the stack follows a container-based architecture using Docker, ensuring portability and ease of deployment. Figure 2 illustrates how the components interact in a simplified way:

1) **API:** serves as the "entry point" of the stack. The client application sends inference requests to the API, which processes or forwards the tasks to other components. When a client sends a request, it immediately receives a *Job ID*, allowing progress tracking without waiting for the processing to finish.

TABLE I
PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS FOR RECORD CLASSIFICATION.

Number of training samples	Neural Network		Decision Tree		Random Forest		SGD Classifier	
	20K	200K	20K	200K	20K	200K	20K	200K
Training duration	*00:00:45*	*00:08:16*	*00:00:34*	*00:21:15*	*00:00:39*	*00:15:57*	*00:01:30*	*01:27:03*
Accuracy (%)	*97.06*	*97.12*	*99.05*	*99.72*	*98.90*	*99.53*	*86.96*	*93.04*
Number of manually tested samples	50,208	50,208	50,208	50,208	50,208	50,208	50,208	50,208
Occurrences of "unknown words"	26,652	13,333	26,652	13,333	26,652	13,333	26,652	13,333
Number of errors (a–e)	*1,474*	*1,444*	*479*	*141*	*550*	*234*	*6,548*	*3,496*
a) Fewer than 100 training samples	*387 / 26.26%*	*107 / 7.41%*	*83 / 17.33%*	*16.31%*	*131 / 23.82%*	*70 / 29.91%*	*749 / 11.44%*	*185 / 5.29%*
b) Unknown word occurrence	*201 / 13.64%*	*685 / 47.44%*	*66 / 13.78%*	*40 / 28.37%*	*51 / 29.27%*	*53 / 22.65%*	*2,398 / 36.62%*	*1,859 / 53.18%*
c) Both (a + b)	*729 / 49.46%*	*99 / 6.86%*	*229 / 47.81%*	*25 / 17.73%*	*268 / 48.73%*	*58 / 24.79%*	*2,174 / 33.20%*	*178 / 5.09%*
d) No samples in training set	*68 / 4.61%*	*3 / 0.21%*	*68 / 14.20%*	*3 / 2.13%*	*68 / 12.36%*	*3 / 1.28%*	*68 / 1.04%*	*3 / 0.09%*
e) Other causes	*89 / 6.04%*	*550 / 38.09%*	*33 / 6.89%*	*50 / 35.46%*	*32 / 5.82%*	*50 / 21.37%*	*1,159 / 17.70%*	*1,271 / 36.36%*

TABLE II
NEURAL NETWORKS USED TO COMPARE PERFORMANCE WITH DECISION TREE AND RANDOM FOREST

Parameter	**Neural Network A**	**Neural Network B**
Number of neurons	10	60
Epochs	20	30
Training duration	∼16 minutes	∼43 minutes
Number of errors	293	88
Accuracy (%)	99.42%	99.82%

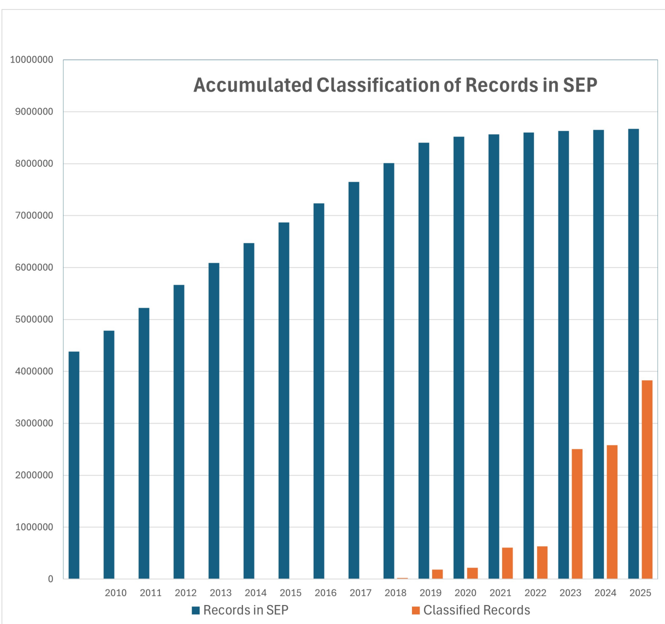

Fig. 1. Accumulated classification of records in SEP system.

2) **Database:** acts as the "central archive" of the stack, storing data used by the models. The API accesses the database whenever information retrieval is required.
3) **Workers:** perform the data processing and model training tasks.
   - **Workers Pub:** process tasks sent to the queue and store results in the database component.

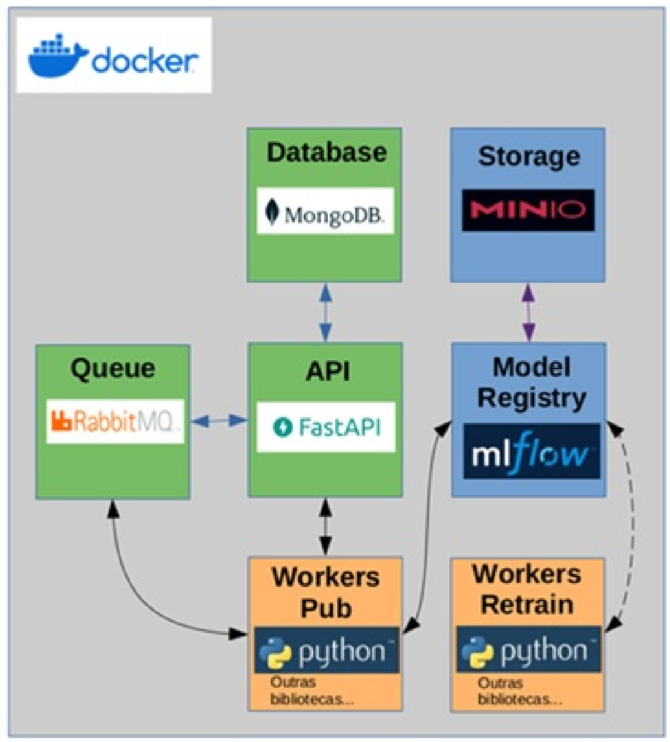

Fig. 2. System architecture of the modular machine learning stack.

   - **Workers Retrain:** retrain models with new data and also save artifacts in storage.
4) **Queue (RabbitMQ):** functions as a "task queue." Requests arrive here and are distributed to the workers, allowing multiple tasks to be processed asynchronously and efficiently.
5) **Storage and Model Registry:** the storage component retains all files and intermediate results, while the model registry maintains versioning and records of trained models. Together, they allow model tracking and reuse without impacting other components.

The communication between components follows a logical flow: the client application sends a request to the API → the API queries the database if necessary → the API sends the

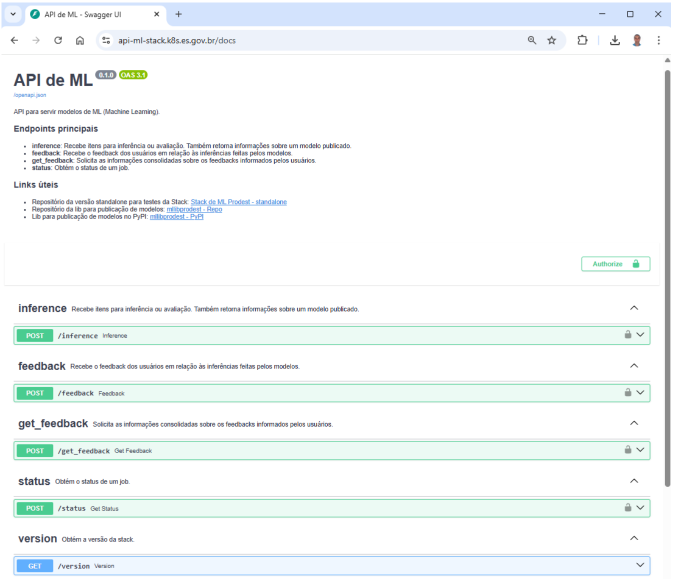

Fig. 3. Swagger UI of the Machine Learning API used by SEP system, showing the available endpoints for inference, feedback, and model management.

task to the queue → the workers process the task → the results are saved in storage and registered in the model registry. This flow ensures that each component has a clear role, simplifying maintenance and scalability.

For classification or regression models to operate within the stack, the use of the **mllibprodest** library is required, as it provides standardized interfaces. This guarantees that any model can be integrated without the need to modify the infrastructure.

This modular and standardized approach allows researchers and developers to focus on designing and testing *machine learning* algorithms, while the stack manages operational aspects such as task queuing, data storage, model versioning, and job status monitoring. The flexibility of the stack makes it suitable for both academic experiments and production environments that demand scalability and reliability.

Figure 3 presents the Swagger UI of the Machine Learning API, which enables the SEP system to interact with the classification models through standardized endpoints.

VII. Discussion

The reported experience shows that the introduction of Artificial Intelligence (AI) into archival processes is not limited to technical efficiency gains but leads to a reconfiguration of archival practice itself.

The application of supervised learning models in the context of the SEP system demonstrated that it is possible to automate the classification of paper-based records while respecting the principles of provenance and organicity highlighted by Foscarini [8], according to which classification should reflect the institutional functions that generate the records rather than their textual content alone. In this sense, the use of neural networks proved to be compatible with the functional approach of contemporary archival science, since the algorithms were trained based on the relationships between records and administrative activities, thus preserving the archival bond between them.

The performance observed can be partially explained by a phenomenon identified in two public agencies that are, in fact, the largest producers of administrative processes in the SEP system. In one of these agencies, an institutional Guideline for Records Creation defined a controlled vocabulary and standardized textual formulations, resulting in highly homogeneous records. In the other, the automation of process creation — through integration between its business system and the Electronic Process System (SEP) — standardized the content automatically inserted into key fields of the records. Although these practices were limited to only two agencies, they generated internally consistent categories that naturally contributed to improved model performance. This contextual information helps explain the accuracy results while remaining fully compatible with the broader scenario of heterogeneous records present in the dataset.

However, the experience also revealed challenges inherent in integrating technology with archival practice. In some cases, the absence of corresponding items in the institutional classification scheme led to prediction errors that initially appeared as model "hallucinations" but actually indicated semantic gaps and inconsistencies within the record management instruments themselves. This finding corroborates Pareja's [10] observation that the adoption of AI in public administration must be accompanied by efforts in semantic harmonization and system integration to ensure interoperability and reliability of results.

Another aspect discussed was the impact of the deactivation of classification schemes and changes in organizational structures, which directly affected the model's performance. This phenomenon illustrates the point raised by Menezes *et al.* [16] regarding the cultural and institutional obstacles that hinder the incorporation of AI in the public sector, such as the lack of standardization, bureaucratic rigidity, and insufficient technological infrastructure. Even so, the combination of human efforts and machine learning proved to be productive: while human classifiers ensure adherence to archival principles, the automated model expands scalability and reduces processing time, forming a decision-support system that evolves toward supervised automation.

Thus, the application of AI to archival classification should be understood as a hybrid process in which neural networks operate in complementarity with archival knowledge. The outcome is not merely a gain in efficiency but a step forward toward a Computational Archival Science, in which artificial intelligence acts as a mediator between the functional logic of archives and the need for scalability and interoperability in systems for managing archival records.

VIII. Conclusion and Future Work

The use of Artificial Intelligence (AI) for the classification of paper-based records has shown potential to accelerate the organization of accumulated collections and contribute to the modernization of governmental records management. The experience with artificial neural networks demonstrated that supervised learning models can achieve high levels of accuracy at low computational cost, provided they are trained with consistent data representative of institutional functions. Thus, AI serves not only as an automation tool but also as a diagnostic instrument for identifying structural weaknesses in archival classification, highlighting the need to revise classification schemes and controlled vocabularies.

Furthermore, the acceleration of the record classification process enabled by Artificial Intelligence made it possible, once classified, to verify the retention periods, even allowing for the elimination of paper-based records. This practice resulted in a significant reduction in costs related to external record storage, contributing to the optimization of archival management within public administration.

In line with Pareja [10], the adoption of AI-based solutions in the public sector should be aligned with digital government strategies, grounded in principles of transparency, efficiency, and ethics. As emphasized by Menezes *et al.* [16], technological integration requires not only infrastructure but also governance policies, staff training, and regulatory frameworks that ensure the responsible use of AI. In this context, the SEP system case can be seen as an example of digital transformation guided by archival values, in which technology is applied to strengthen the authenticity and reliability of records.

For future work, the plan is to integrate the classifier into digital archival management systems, enabling the automated processing of both legacy records and newly created digital records. Finally, within this context, this study reinforces that artificial intelligence, when developed in dialogue with archival theory, has the potential to transform archival management and support the preservation of institutional memory, contributing to the consolidation of effective records management practices.

References

[1] L. Duranti and P. C. Franks, Encyclopedia of Archival Science. *Lanham, MD: Rowman & Littlefield*, 2015.

[2] J. Teixeira, J. O. Almeida, T. B. S. Gava, T. Canelhas, and J. M. M. Dorigueto, "Artificial intelligence applied to paper support documents: The case of the SEP system in ES". Archeion Online, Special Issue – X CNA, pp. 95–105, Jan./Mar. 2025. [In Portuguese].

[3] T. R. Schellenberg, Modern Archives: Principles and Techniques. Chicago, IL: University of Chicago Press, 1956.

[4] M. Duchein, "Theoretical principles and practical problems of respect des fonds in Archival Science". Archivaria, no. 16, pp. 64–82, 1983.

[5] L. Duranti, "Diplomatics: new uses for an old Science". Lanham, MD: Scarecrow Press, 1998.

[6] S. Russell and P. Norvig, "Artificial intelligence: a modern approach". 4th ed. Hoboken, NJ: Pearson, 2021.

[7] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. The MIT Press, 2016, 800 pp, ISBN: 0262035618.

[8] F.Foscarini, Records classification and functions: An archival perspective, Journal of the American Society for Information Science and Technology, vol.57, no.11, pp.1442–1452, 2006.

[9] R.Marciano, V.Lemieux, M.Hedges, M.Esteva, W.Underwood, M.Kurtz, and M.Conrad, Archival Records and Training in the Age of Big Data, in *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education (Advances in Librarianship, Vol.44B)*, J.Percell, L.C.Sarin, P.T.Jaeger, and J.C.Bertot, Eds. Bingley, UK: Emerald Publishing, 2018, pp.179–199.

[10] H.P.Pareja, Artificial Intelligence as a Driver of Digital Government Transformation: Applications and Trends in Government Operations in the Philippines, *Policy and Governance Review*, vol.9, no.3, pp.302–317, 2025, doi: 10.30589/pgr.v9i3.1257.

[11] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge University Press, 2008.

[12] S. Ruder, "An overview of gradient descent optimization algorithms", arXiv preprint arXiv:1609.04747, 2016.

[13] F. Chollet, Deep Learning with Python. Shelter Island, NY, USA: Manning Publications, 2018.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," Journal of Machine Learning Research, vol. 15, pp. 1929–1958, 2014.

[16] C. I. C. Menezes, D. A. Machado, G. A. A. Simonini, and V. O. B. U. Fernandes, "Artificial intelligence and the challenges for the political-administrative organization of the state," Caderno da Escola Superior de Gestão Pública, Política, Jurídica e Segurança, vol. 7, no. 1, pp. 81–92, 2024. [In Portuguese].